Douban Book Review Classification based on CNNs

周姣美

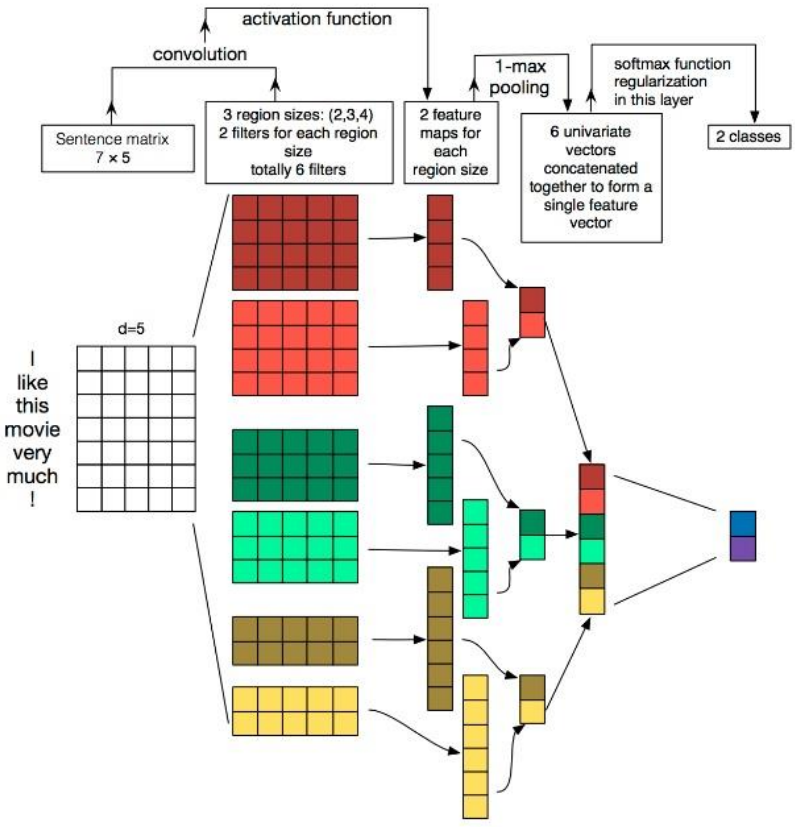
Background

文本分类是自然语言处理领域中的一个经典问题，最早主要是采用规则（pattern）进行分类，但是效率和准确率都较为有限；随后，统计学习方法得到了广泛的应用，人们逐渐采用特征工程+机器学习方法对文本进行分类，其中特征工程分为文本预处理、特征提取和文本表示等几个部分，将自然语言中的文本转化成计算机可以处理的形式。

近年来，深度学习方法在自然语言处理领域取得巨大成功，不同于此前反复的人工特征提取的过程，利用深度学习方不仅法可以解决文本表示高纬度高稀疏的问题，用卷积神经网络（CNN）等网络结构还可以抓取文本中的特征，从而实现文本的自动分类。

Method

本文使用的模型是经典的 CNN 模型，来自论文：Convolutional Neural Networks for Sentence Classification，其原理如下图所示：



第一层是 embedding 层，获取原始文本的句子矩阵，将词/字作为向量输入；第二层则是卷积层，第三层是池化层，不同长度的句子经过池化之后都表示为相同长度的向量，最后是 softmax 层，输出两个类别的概率。

Data

本文使用的数据是豆瓣图书评论数据，共 39764 条，分为积极和消极两类，分别标注为 P 和 N，从下表中可以看出，两类数据比较均衡。并且，观察原始文本可以看出，大部分为短文本，且不同于报纸或文学作品，更接近微博等较为生活化、口语化的语言，因为本文采用的预训练词向量为微博文本。

按照 6：2：2 的比例将原始数据划分训练集、验证集和测试集。

训练集	验证集	测试集		P	N
23858	7953	7953		19891	19873
60.00%	**20.00%**	**20.00%**		**50.02%**	**49.98%**

Result and analysis

训练 CNN 模型时使用的参数为：dropout = 0.6，epochs = 50，learning_rate = 1e-4，pad_size = 50，batch_size = 25。

以下是模型在测试集上的表现：

Test Loss: 0.2, Test Acc: 92.59%

Precision, Recall and F1-Score...

		precision	recall	f1-score	support
N	0	0.9268	0.9242	0.9255	3960
P	1	0.9251	0.9276	0.9263	3993
accuracy				0.9259	7953
macro avg		0.9259	0.9259	0.9259	7953
weighted avg		0.9259	0.9259	0.9259	7953

Time: 0:29:42

可以看到，准确率、召回率和 F1 值都在 92%左右，说明卷积神经网络对于较短文本的二分类效果还是很好的。此外，本文还使用了一些经典的机器学习模型进行分类，从而可与 CNN 模型进行比较。文本表示采用 TFIDF 方法，数据则是将前文中的训练集和验证集合并作为训练集进行训练，测试集则与 CNN 模型完全一致，均为默认参数。

最终得到的 F1 值和训练时间如下表所示：

	BernoulliNB	MultinomialNB	linearSVC	KNN	DecisionTree	RandomForest	LogisticRegression	CNN
training time(秒)	0.013	0.007	0.148	0.005	13.784	23.298	1.337	29分42秒
F1	0.904	0.922	**0.932**	0.533	0.893	0.897	**0.927**	**0.926**

通过横向比较，可以看到支持向量机和逻辑回归的效果均优于 CNN，且运行时间很少。分析可能有以下原因：1.训练集过拟合，到第 50 轮时，如下图所示，CNN 在训练集上的准确率基本已经差不多是 100%；2.模型还需要进一步调整参数，例如 batch_size（批尺寸）是否没有调整到最适合此数据集的大小；3.本文采用的 CNN 模型是比较早将卷积神经网络用于文本分类的模型，其结构较为简单，近年来也出现了一些改进的模型，可能需要用更新效果更好的 CNNs。